DOCUMENT RESUME

ED 400 295 TM 025 577

AUTHOR Hennings, Sara S.; And Others
TITLE A Comparison of Equating Methods Applied to Performance-Based Assessments.
PUB DATE Apr 96
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classification; Comparative Analysis; *Cutting Scores; Elementary School Students; Elementary Secondary Education; *Equated Scores; High School Students; *Item Response Theory; Norm Referenced Tests; *Performance Based Assessment; *Raw Scores
IDENTIFIERS Anchor Tests; *Equipercentile Equating; *Graded Response Model; Partial Credit Model

ABSTRACT

Three methods of equating performance assessments are compared. Equipercentile equating, item response theory (IRT) partial credit model, and IRT graded response model methods were applied to the same data using an anchor test design. Comparisons are based primarily on the resulting raw score-to-raw score tables and the effects these tables have on the classification of students in terms of two cut scores. The design entailed administering the performance assessment and a norm-referenced test, the anchor test, to each student. Matched data from between 4,950 and 7,820 students were analyzed at grades 3, 5, 8, and 10 in 1993 and 1994. The norm-referenced tests were the Iowa Tests of Basic Skills at the lower grades, and the Tests of Achievement and Proficiency for grade 10. All three methods produced relatively consistent results. In examining the percentage of students classified into performance levels in 1993 versus 1994 with the three equating methods, the equipercentile method produced the most consistent results. Both the partial credit and graded response methods had more deviation in the classification rates with the graded response model being more inconsistent. (Contains two figures, three tables, and seven references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
SARA S. HENNINGS

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 295

A Comparison of Equating Methods Applied to Performance-Based Assessments

Sara S. Hennings
Thomas M. Hirsch
The Riverside Publishing Company

Liru Zhang
Delaware Department of Public Instruction

BEST COPY AVAILABLE

Paper presented at the Annual Meeting of the National Council on Measurement in Education
New York City, NY
April, 1996

TM025577

A Comparison of Equating Methods Applied to Performance-Based Assessments

Introduction

Many statewide assessment programs now include the administration of performance assessments. The results from these performance assessments are often reported in terms of the proportion of students classified as meeting some performance standard or goal with the intention of then comparing these results from year to year. The transparent nature of performance assessments necessitates the use of alternate forms in subsequent years. Typically the standard is set on the performance assessment administered the first, or baseline, year. Thus, a comparable cut score must be determined for the alternate forms to permit the desired longitudinal comparison. The equating of alternate test forms of multiple-choice items has been studied in detail by researchers and implemented by practitioners on a regular basis. However, the equating of tests with open-ended items has not been examined in the same detail. This is especially true for equating methods using polytomous item response theory models.

The purpose of this study is to examine the issue of proper methods for equating tests with open-ended items. This study compares three methods of equating performance assessments: equipercentile, IRT Partial Credit Model, and IRT Graded Response Model. All methods were applied to the same data using an anchor-test design. Comparisons among the results are based primarily on the resulting raw score–to–raw score tables, and the effect these tables have on the classification of students in terms of two cut scores.

Background

There are a large number of published research studies that pertain to equating methods applied to multiple-choice tests. Using these same procedures on performance assessments seems logical. However, few studies have examined equating methods applied to performance assessments. A study by Huynh and Ferrara (1994) used equipercentile and the Partial Credit Model to equate a performance assessment. Using an equivalent-groups design, this study showed that the two methods do not give equivalent results when score distributions are skewed. For assessments that are composed of moderately easy items, the Partial Credit Model may be more useful.

A study by Livingston (1994), which is similar to the design employed in this study, examined using a multiple-choice test as an anchor. In Livingston's study several equatings using a multiple-choice test as an anchor were compared with the equating of two half-tests using equivalent groups. He found that the anchor test equating was clearly preferable to an assumption of equal difficulty of the half-tests.

Many studies have compared the results of equating tests consisting of multiple-choice items using equipercentile, Rasch, and three-parameter IRT methods. This study takes that body of research and expands it into the area of performance-based assessments.

Methods

Instruments Used. In the spring of 1993, performance-based assessments in mathematics and reading were first administered to all Delaware students in grades 3, 5, 8, and 10. In addition, the norm-referenced *Iowa Tests of Basic Skills®* (*ITBS®*), Form K, Survey Battery was administered at grades 3, 5, and 8, and in grade 10 the *Tests of Achievement and Proficiency™* (*TAP®*), Form K, Survey Battery was administered.

Using these data in a standard-setting process, standards for the performance assessments were established creating two cut points. This allowed for the classification of student performance into three categories: 1) meets or exceeds the standard, 2) approaches the standard, and 3) considerably below the standard. To maintain consistency in the standards across adjacent years, a relationship between the raw scores on the 1993 test forms and the 1994 test forms was determined by equating the forms. Using this relationship, cut scores on the 1994 test form were found which were comparable to the cut scores on the 1993 test form at each grade level. This allowed for the classification of student performance in 1994 using standards comparable to those applied in 1993.

Equating Methods. In this study, the anchor-test design entailed administering the performance assessment and the norm-referenced test—the anchor test—to each student. Matched data from between 4950 to 7820 students were analyzed at each grade. In 1993, Delaware students were administered Form A of the performance assessments in both mathematics and reading as well as the *ITBS/TAP*. In 1994, Delaware students were administered Form B of the performance assessments and the same form of the

ITBS/TAP. This allows the use of the *ITBS/TAP* as an anchor test in equating Form A and Form B. Summary statistics for the 1993 data can be found in table 1. Table 2 shows the summary statistics for the 1994 data.

Table 1

Descriptive Statistics for the 93 Performance Assessment Raw Scores and the NRT Scale Scores

Mathematics

	Performance Assessment Raw Score				NRT Scale Scores		
Grade	# Items	Maximum Score	Mean	SD	Mean	SD	N
3	22	34	23.76	5.81	181.76	21.71	6031
5	15	27	12.42	5.85	213.74	27.24	7093
8	19	34	17.34	7.80	251.86	34.17	6556
10	15	25	16.22	4.75	270.88	36.40	4953

Reading

	Performance Assessment Raw Score				NRT Scale Scores		
Grade	# Items	Maximum Score	Mean	SD	Mean	SD	N
3	18	25	18.90	3.63	183.51	21.83	6027
5	16	20	15.20	2.50	213.44	25.37	7099
8	18	25	18.16	4.39	250.05	32.61	6574
10	24	29	22.23	5.42	274.30	36.66	5023

Table 2

Descriptive Statistics for the 94 Performance Assessment Raw Scores and the NRT Scale Scores

Mathematics

	Performance Assessment Raw Score				NRT Scale Scores		
Grade	# Items	Maximum Score	Mean	SD	Mean	SD	N
3	21	33	21.55	6.03	181.12	21.42	7627
5	14	23	11.01	5.73	211.19	27.91	7493
8	18	30	9.94	7.49	249.42	35.05	7138
10	12	27	11.57	6.43	268.67	37.65	5699

Table 2 (continued)

Reading

	Performance Assessment Raw Score				NRT Scale Scores		
Grade	# Items	Maximum Score	Mean	SD	Mean	SD	N
3	12	25	18.59	4.28	181.02	21.57	7819
5	15	25	17.37	4.74	210.49	25.72	7696
8	13	26	20.24	4.22	248.44	32.94	7178
10	11	25	19.21	4.58	271.61	37.61	5739

Equipercentile Equating. In equipercentile equating, two scores are considered equivalent when they correspond to the same percentile rank. The raw scores on the 1993 performance assessments (Form A) were equated to the scale score of the norm-referenced test (NRT). Separately, the raw scores on the 1994 performance assessments (Form B) were equated to the scale scores of the NRT. Using the NRT as the anchor scale provided the link between the raw scores on Form A and Form B.

Partial Credit Model (PCM). The model used in this equating design was the Partial Credit Model (Wright & Masters, 1982). The PCM is an extension of the Rasch model that allows for the calibration of polytomous items (items not scored as simply right or wrong). Use of this model produces a single raw score to ability relationship. This is due primarily to the fact that the use of this model assumes a constant discrimination for all test items similar to the Rasch model assumption; therefore, ability estimates are not dependent on "pattern scoring" as they would if an IRT model which recognizes different discrimination in items were used. Actually, the Rasch Model can be considered a special type of PCM.

All item estimation was performed using the computer program BIGSTEPS (Linacre & Wright, 1994). The equating involved the following steps.

1) Analyze the student responses for both Form A of the performance assessment and the *ITBS/TAP* in a single computer run.

2) Analyze both Form B and the *ITBS/TAP* in a single run fixing the item parameters for the *ITBS/TAP* to those found in step 1. This places the item parameter estimates for Form B onto the same scale as Form A.

3) Using the item parameters derived in step 1 for Form A only, compute a raw score to ability table.

4) Using the item parameters derived in step 2 for Form B only, compute a raw score to ability table.

5) Because the abilities found in both tables derived from steps 3 and 4 are on the same scale, this relationship can be used to determine the raw score–to–raw score relationship between Form A and Form B.

The Graded Response Model (GRM). The Graded Response Model (Samejima, 1969) is similar to the Partial Credit Model in that it allows for the calibration of items which are not scored dichotomously. The major difference between the models is the inclusion of a discrimination parameter in GRM. The GRM contains a single discrimination parameter for each item as well as a difficulty parameter for each item category (i.e., score points). The inclusion of the discrimination parameter results in "pattern scoring" for the assignment of abilities to students and does not produce a single raw score to ability relationship. For equating purposes this pattern scoring was not utilized and instead the test characteristic curve was used to estimate a raw score to ability relationship for the two tests being equated.

The procedure used with this model involved the creation of a data set consisting of both 1993 and 1994 data. To the data from 1993, a string of 9's signifying not administered was added to represent the 1994 performance assessment. Similarly, a string of 9's to signify not administered were attached to the 1994 data filling the slots for the 1993 data. The responses to the *ITBS/TAP* were common to all lines in the data set, thus serving as the anchor. This data set allowed for the simultaneous calibration of the 1993 and 1994 performance assessments with the *ITBS/TAP* serving as the anchor test. All calibrations were completed using MULTILOG (1991) and the Graded Response Model (Samejema, 1969). After the calibration were completed, the item parameters for Form A and Form B of the performance assessments were on the same scale. The equatings were then completed similarly to that described for the Partial Credit Model:

1) Using the item parameters for Form A only, compute a raw score to ability table.

2) Using the item parameters for Form B only, compute a raw score to ability table.

3) Because the abilities found in both tables derived from steps 1 and 2 are on the same scale, this relationship can be used to determine the raw score–to–raw score relationship between Form A and Form B.

Due to MULTILOG limitations, the analyses using the GRM were based on only 500 students per equating.

The three equatings were performed on the mathematics and reading performance assessments for grades 3, 5, 8, and 10. Comparisons among the raw score–to–raw score tables resulting from the three equating methods were made. The impact that the equated scores had on the cut scores was summarized in terms of the percent of students falling in each performance category.

Results/Conclusions

Generally across all equatings, the equipercentile and PCM equatings produced very similar results. The equatings using GRM were somewhat more variable relative to the other equatings. This may have been due to the smaller number of students used for the calibrations. The graphic representation of the resulting raw-score to raw-score relationships are illustrated in figures 1 and 2.

In the equatings of the mathematics performance assessments in grades 8 and 10 all three methods produced similar results. For grade 3 the GRM is the most variable and for grade 5 the PCM is the most variable relative to the other two equating methods. For reading all three methods produced similar results for grades 3 and 10 in the raw score range surrounding the cut scores. In grades 5 and 8 the GRM tended to vary more than the other two methods.

Simply producing different equating results does not imply that a particular method is inaccurate relative to the other methods. However, in this study it is possible to examine the effects of the equating results in terms of the percent of students being classified into the three performance levels. Given the size of the state testing program, it is typically assumed that there would be small changes in the classification rates from one year to the other. That is, one would not expect extreme increases (or decreases) in performance

BEST COPY AVAILABLE

at the state level, therefore, it should be possible to use this criterion for comparison among the three equating methods. The percent of students being classified into the three performance levels by grade level are presented in table 3.

Table 3

Comparison of the Percent of Students in Each Performance Category for Each Equating Method

Grade	Performance Standards	Mathematics				Reading			
		93	94			93	94		
			EQPR	PCM	GRM		EQPR	PCM	GRM
3	Meets/Exceeds the Standard	18.2	17.7	17.7	17.7	13.5	15.7	7.2	15.7
	Approaches the Standard	46.7	47.5	42.0	57.3	57.7	52.2	60.7	58.5
	Below the Standard	35.1	34.8	40.3	25.1	28.8	32.1	32.1	25.7
5	Meets/Exceeds the Standard	13.4	12.3	20.9	8.4	16.5	11.7	11.7	11.7
	Approaches the Standard	40.0	39.1	30.5	48.9	70.5	58.3	68.4	80.9
	Below the Standard	46.5	48.6	48.6	42.6	13.0	30.0	19.9	7.4
8	Meets/Exceeds the Standard	13.0	12.0	12.0	16.2	12.8	20.0	8.9	8.9
	Approaches the Standard	43.3	41.4	36.3	37.2	64.1	55.9	67.0	76.3
	Below the Standard	43.7	46.6	51.6	46.6	23.1	24.2	24.2	14.8
10	Meets/Exceeds the Standard	11.9	13.5	10.6	10.6	22.9	27.8	27.8	15.6
	Approaches the Standard	54.9	51.1	54.1	54.1	56.3	47.6	47.6	59.7
	Below the Standard	33.2	35.4	35.4	35.4	20.8	24.7	24.7	24.7

EQPR = Equipercentile Equating
PCM = Partial Credit Model
GRM = Graded Response Model

Overall it appears that the equipercentile and the PCM equatings produced the most consistent equatings. As suspected after the discussion of the graphs, the GRM tended to produce the most variable results.

For mathematics grade 10, all three of the methods produced similar results based on the percent of students being classified as "Meets/Exceeds the Standard." The equipercentile method produced 1994 results most similar to 1993 in grades 3, 5, and 8. The PCM method produced the most discrepant results in grade 5 and the GRM produced discrepant results in grades 5 and 8.

Examining the percent of students being classified as "Meets/Exceeds the Standard" for reading, the equipercentile method produced 1994 results most similar to 1993 in all grades except grade 8. The PCM

equatings produced consistent results in all but grade 3, and the GRM had consistent results in all grades except grade 10. These results illustrate the agreement between the three methods.

It is important to remember when interpreting these results that the distributions for these performance assessments are not as "smooth" as is normally found with a norm-referenced test. Therefore relatively large differences in the percent classified in a performance level may be the result of a single raw score point difference at the cut score. Further, this raw-score difference may be caused by a small fractional difference in the 1993 to 1994 raw-score to raw-score relationship.

Educational Importance

With the increased use of performance assessments in statewide assessment programs, the need for effective equating methods and designs involving these assessments is imperative. Although test users appreciate the use of alternative item formats, they still insist on the same high degree of psychometric quality found in most commercially produced NRTs. This includes the ability to equate alternate test forms allowing for comparisons in students' performance based on different forms. This study compared three equating methods that can be used with performance assessments: equipercentile, Partial Credit Model, and Graded Response Model.

The results from this study indicate that all three methods produce relatively consistent results. In examining the percent of students classified into the performance levels in 1993 vs. 1994 with the three equating methods, the equipercentile equating method produced the most consistent results. Both the PCM and GRM methods had more deviation in the classification rates with the GRM being most inconsistent.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Huynh, H. & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. *Journal of Educational Measurement*, 31, 125-141.

Linacre, J. M. & Wright, B. D. (1994). BIGSTEPS [Computer program]. Chicago, IL: MESA Press.

Livingston, S. A. (1994). *Equating constructed-response tests through a multiple-choice anchor: A small-scale empirical study.* Unpublished statistical report, Educational Testing Service, Princeton, NJ.

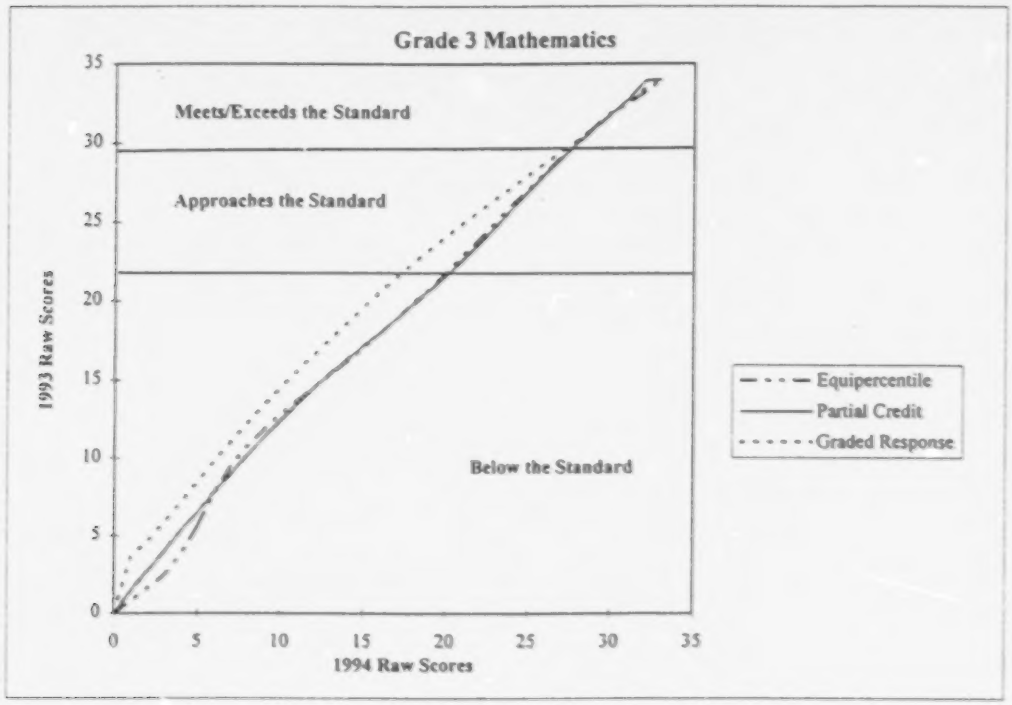
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34 (Monograph Number 17).

Thissen, D. (1991). MULTILOG [Computer program]. Chicago, IL: Scientific Software International.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Figure 1

Comparison of the 94 Raw Scores Equated to the 93 Raw Scores Using the Three Equating Methods



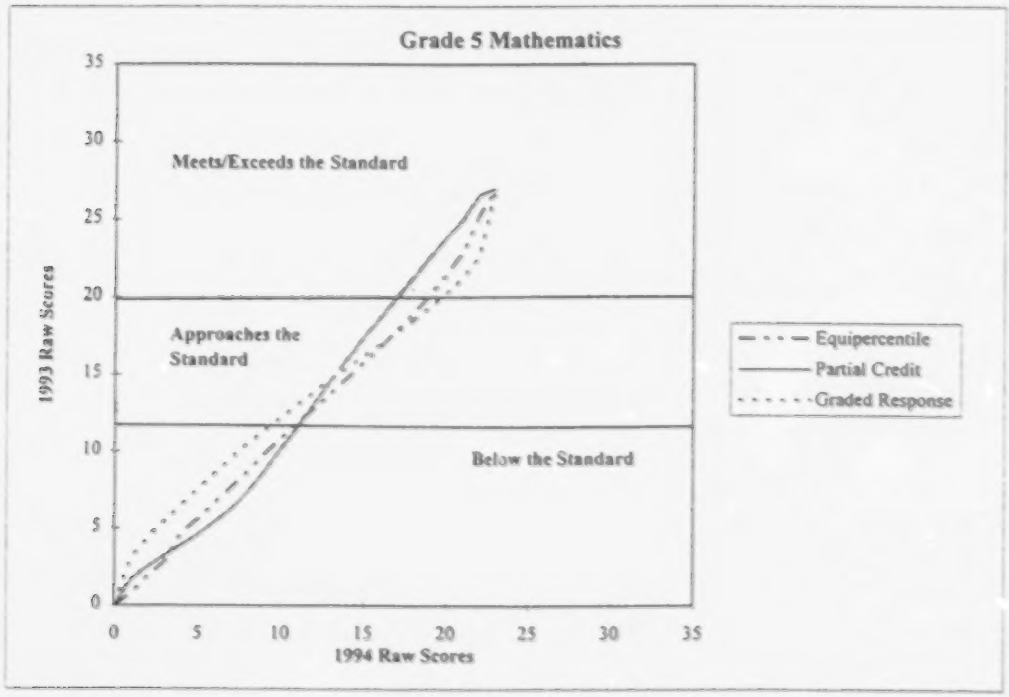
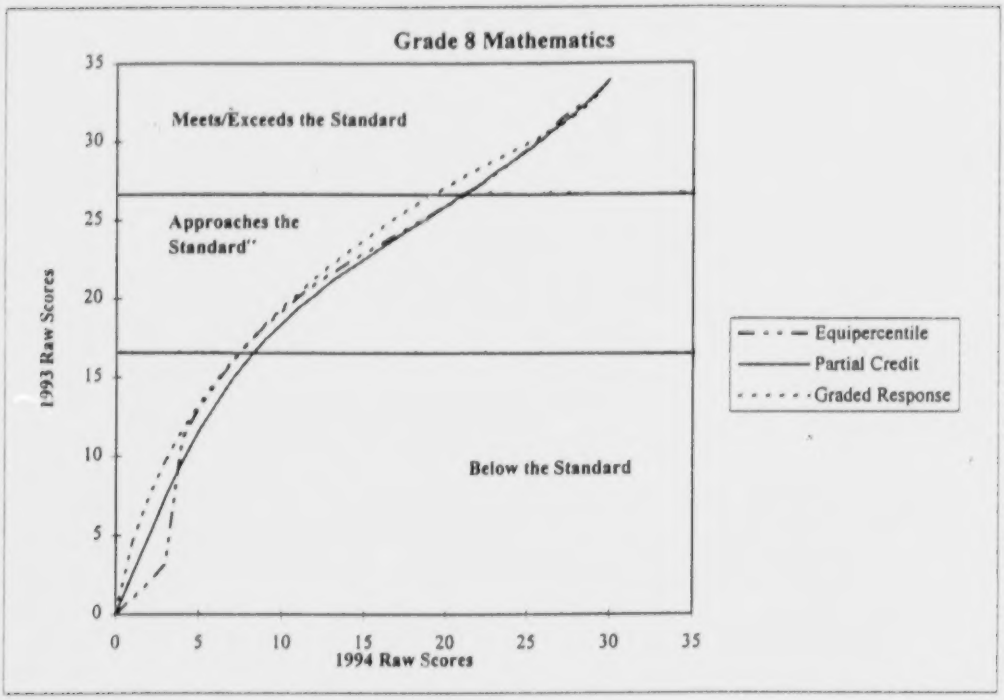


Figure 1 (continued)



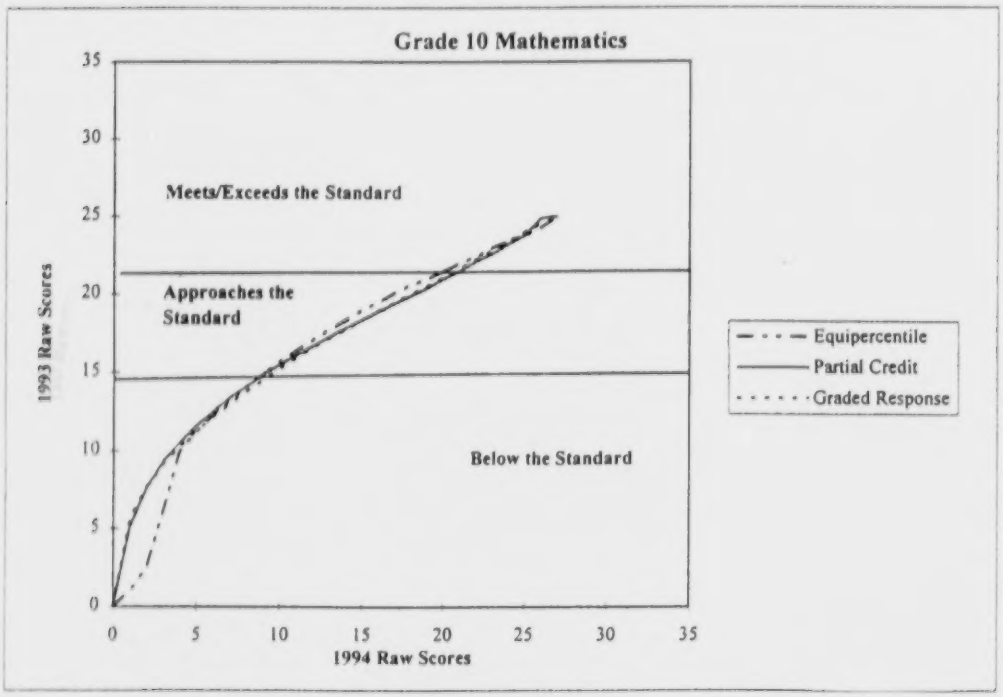
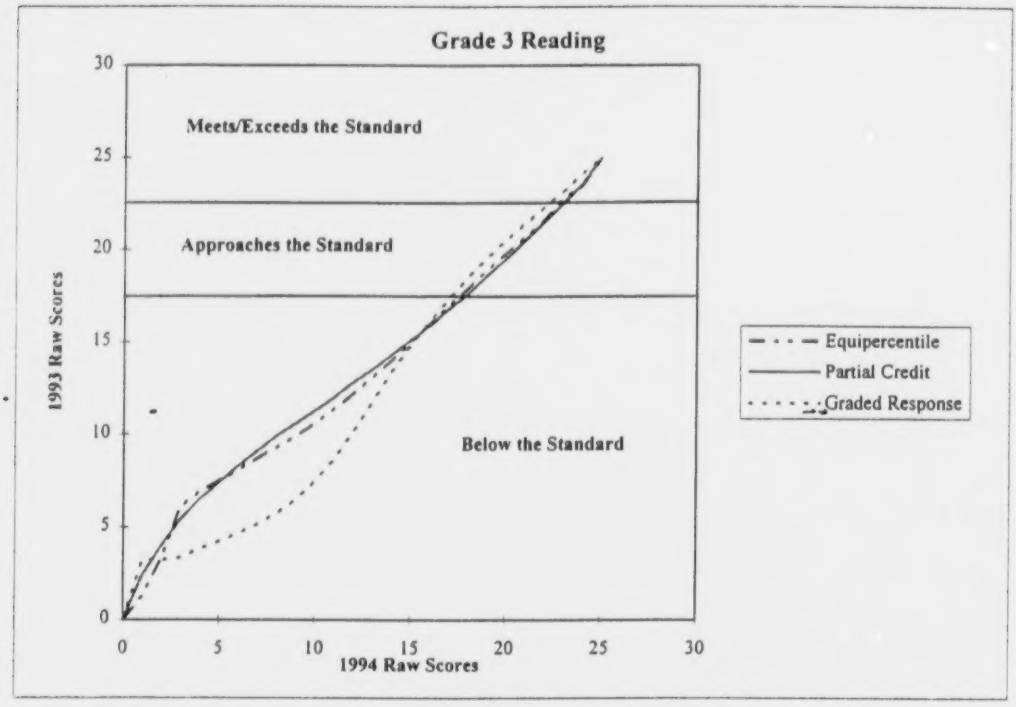


Figure 2

Comparison of the 94 Raw Scores Equated to the 93 Raw Scores Using the Three Equating Methods



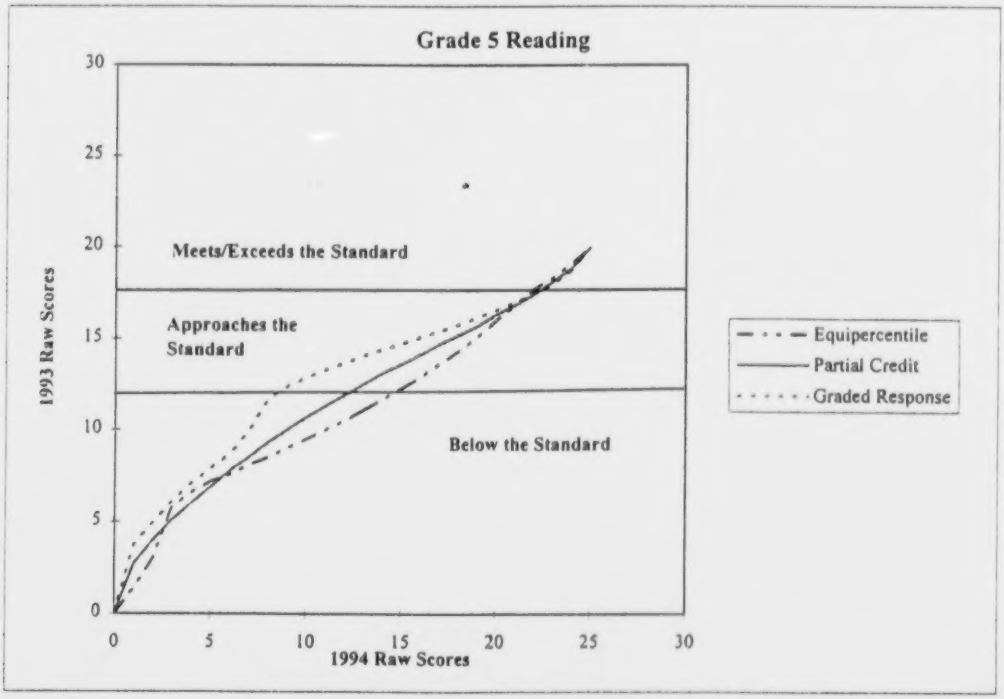


Figure 2 (continued)